Supporting Information for "Application of the AI2 Climate Emulator to E3SMv2's global atmosphere model, with a focus on precipitation fidelity"

James P. C. Duncan[1], Elynn Wu[2], Jean-Christophe Golaz[3], Peter M. Caldwell[3], Oliver Watt-Meyer [2], Spencer K. Clark[2], Jeremy McGibbon[2], Gideon Dresdner[2], Karthik Kashinath[4], Boris Bonev[4], Michael S. Pritchard[4], and Christopher S. Bretherton[2]

[1]University of California, Berkeley

[2]Allen Institute for Artificial Intelligence (AI2), Seattle, USA

[3]Lawrence Livermore National Laboratory, Livermore, California, USA

[4]NVIDIA, Santa Clara, California, USA

Contents of this file

Introduction

In this Supporting Information, we give additional metrics related to ACE's climatological skill and supplementary figures which provide additional perspectives on the figures of the main text. We also provide further details on the computational efficiency of ACE, the vertical coarsening of raw EAMv2 simulations outputs, and the optimization hyperparameters employed during ACE training.

Text S1. Another perspective on ACE's emulation biases

Figure S1 compares ACE's emulation biases to EAMv2's internal variability. The left column labeled "EAMv2 reference vs. EAMv2" displays the bias patterns observed when comparing EAMv2 to itself, which serves as an 'oracle' emulator with the highest climate skill possible in terms of faithfulness to the original simulation, given natural variability due to weather fluctuations. These biases are computed by comparing the unseen reference set, years 64–73 of the EAMv2 simulation run, against the validation target years 54–63. The column labeled "ACE vs. EAMv2" visualizes the same data as the right column of Figure 3 of the main text. Table S1 provides additional bias and RMSE metrics for these variables when evaluating ACE and EAMv2 internally (i.e., against EAMv2 simulation outputs) as in Figure S1 and against historical observations as in the left column of Figure 3.

Text S2. Computational efficiency of ACE

We carried out the 73 year EAMv2 simulation on the Chrysalis supercomputer at Argonne National Laboratory, which is a dedicated E3SM machine[1]. Using 30 CPU nodes on Chrysalis, each of which has 2× 32-core AMD EPYC 7532 CPUs, the simulation achieved 24 simulated years per day, or about 10 seconds per simulation day. After training, we ran

ACE inference using a single NVIDIA A100 40 GB GPU on Lawrence Berkeley National Laboratory's Perlmutter supercomputer with a wall clock time of 1 second per simulation day, an approximate 10x speedup. The discrepancy with the 100x speedup found in Watt-Meyer et al. (2023) is explained by the much larger number of cores used for the EAMv2 simulation compared to the FV3GFS simulation, which used a total of 96 cores across two higher-efficiency AMD EPYC 7H12 CPUs. We estimate the energy consumption of 1 second on 1 A100 GPU at maximum power consumption of 400 W is 0.4 kJ, while 10 seconds on 60 total EPYC 7532 CPUs at 200 W is approximately 120 kJ. This amounts to an approximate 300x energy savings when using ACE as a surrogate for EAMv2.

References

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. doi: 10.48550/arXiv.1412.6980

Kucharski, F., Molteni, F., King, M. P., Farneti, R., Kang, I.-S., & Feudale, L. (2013). On the need of intermediate complexity general circulation models: A "SPEEDY" example. *Bulletin of the American Meteorological Society*, *94*, 25–30. doi: 10.1175/BAMS-D-11-00238.1

Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S. K., Henn, B., Duncan, J., ... Bretherton, C. S. (2023). ACE: A fast, skillful learned global atmospheric model for climate prediction. doi: 10.48550/arXiv.2310.02074

Notes

1. https://climatemodeling.science.energy.gov/news/chrysalis-ready-emerge-e3sm-v2-runs

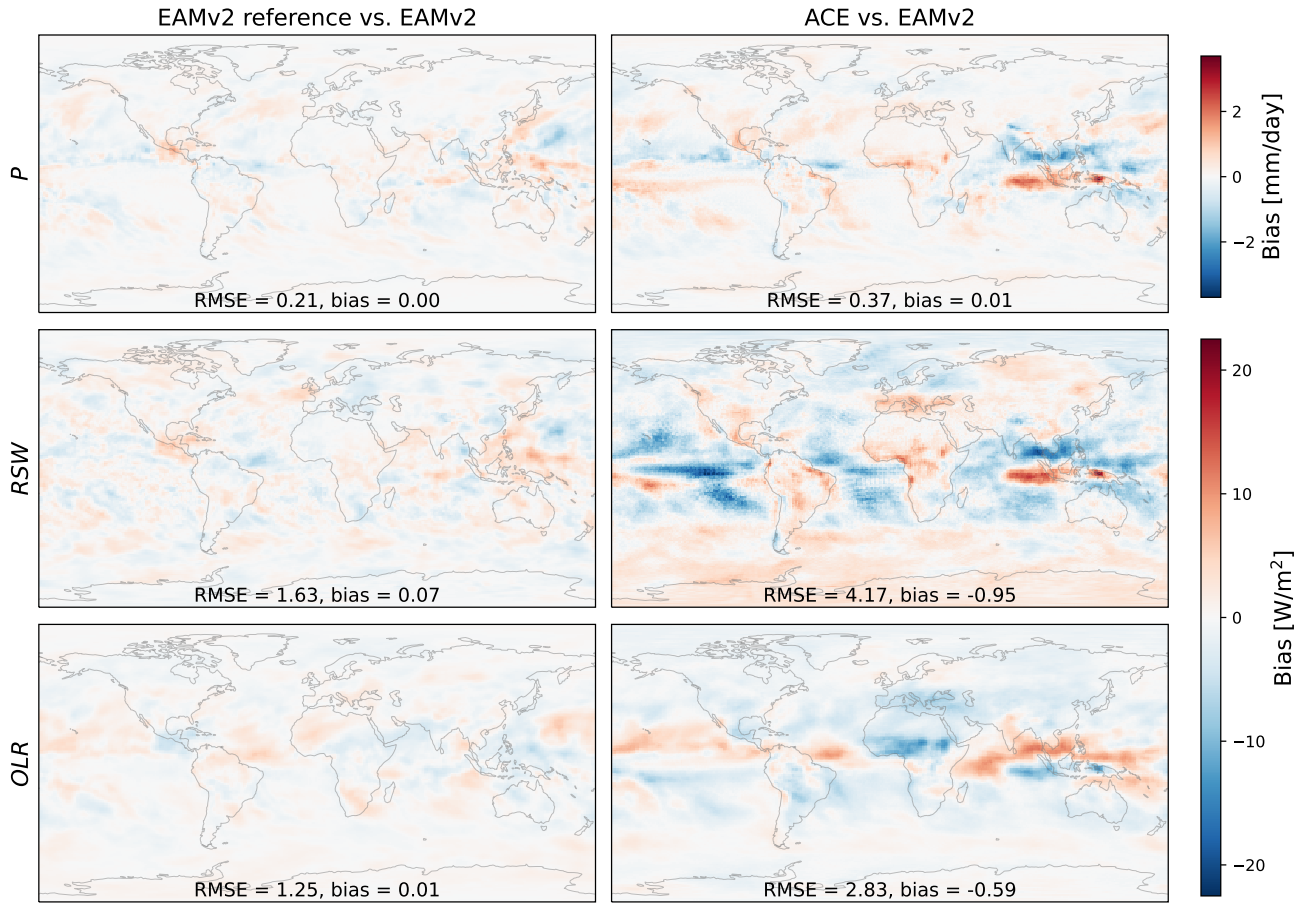

Figure S1. Time average biases (*predicted* - *target*) for precipitation (top row) and top-of-atmosphere outgoing shortwave (RSW, middle row) and longwave (OLR, bottom row) radiative fluxes. The right column ("ACE vs. EAMv2") shows the mean spatial distribution of ACE biases, comparing the generated 6-hourly outputs to the corresponding targets for the same timestep. The left column ("EAMv2 reference vs. EAMv2") compares EAMv2 to itself by recalculating biases using the final 10 years of the simulation set in the place of the *predicted* data, giving a best-case scenario reference.

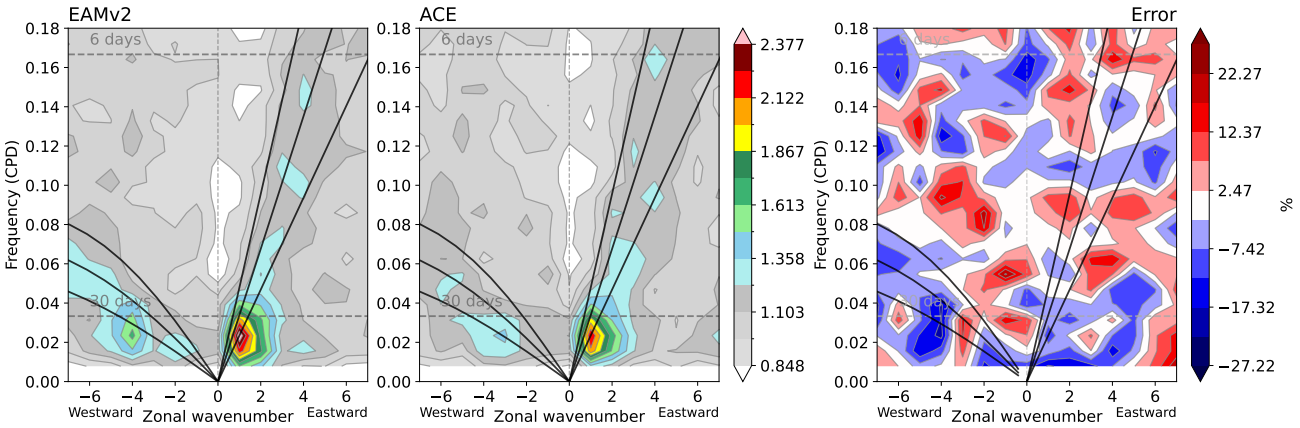

Figure S2. Same as Figure 4 of the main text but zoomed in for a closer look at the tropical spectra between wavenumbers -6 and 6 and frequencies smaller than 0.18. In addition, the third panel displays relative errors within this region, calculated as: $100 \times \frac{\text{predicted power} - \text{target power}}{\text{target power}}\%$.

Table S1. ACE and E3SMv2 biases and RMSEs with respect to various references. ACE_{int}: ACE compared against EAMv2 outputs over the 10 year validation period. EAMv2_{int}: EAMv2 outputs over the 10 year reference period compared against EAMv2 outputs over the 10 year validation period. ACE_{obs}: ACE compared against historical observations. EAMv2_{obs}: EAMv2 outputs over the 10 year validation period compared against historical observations.

	ACE_{int}		EAMv2_{int}		ACE_{obs}		EAMv2_{obs}	
Variable	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
P [mm/day]	5.7e-3	0.37	1.6e-3	0.21	0.20	0.93	0.20	0.96
RSW [W/m^2]	-0.95	4.17	6.7e-2	1.63	-0.38	8.87	0.57	9.19
OLR [W/m^2]	-0.59	2.83	8.5e-3	1.25	-0.77	5.64	-0.17	5.09

Table S2. EAMv2 vertical interface coordinates that were used for vertical coarsening of the raw 3D outputs, reducing the number of vertical levels from 72 to 8 for computational tractability. As in Watt-Meyer et al. (2023), we chose the 9 vertical interfaces listed below that best align with those of the SPEEDY model (Kucharski et al., 2013), in sigma coordinates, assuming a constant reference surface pressure of $p_8^{ref} = 1000$ hPa. The coarsened interfaces are indexed starting from the top of the atmosphere by k from 0 to 8, while the corresponding original EAMv2 interfaces are indexed by I_k. In each grid column, the terrain-following interfacial pressures $p_k = a_k + b_k p_s$ are computed from the hybrid coordinates a_k and b_k and the surface pressure p_s. The original model levels are vertically integrated by mass in order to preserve the total dry air and moisture budget, using the true surface pressure at each point in space and time. For further details, see Watt-Meyer et al. (2023).

k	a_k [Pa]	b_k [unitless]	I_k	p_k^{ref} [hPa]
0	10.0	0.0	0	0.1
1	4943.694	0.0	19	49.4
2	13913.118	0.0	30	139
3	16254.503	0.10464	38	267
4	12435.282	0.31152	44	436
5	8945.939	0.50053	48	590
6	5115.018	0.70804	53	759
7	2027.536	0.87529	61	896
8	0.0	1.0	72	1000

Table S3. Following Watt-Meyer et al. (2023), we employ the Adam optimizer (Kingma & Ba, 2017) with a cosine annealing learning rate schedule decaying to zero by the end of training and use an exponential moving average of the model parameters across training steps. We conducted a thorough hyperparameter search across 29 combinations of batch size, initial learning rate, and number of epochs, arriving at the final choice of hyperparameters based upon a comparison of 10-year time-mean validation metrics, multiyear stability, and visual artifacts. See Watt-Meyer et al. (2023) for additional details on training and SFNO architectural hyperparameters.

Name	Value
Initial learning rate	3×10^{-4}
Number of epochs	50
Batch size	8